CJSS: CANDIDATE-DYNAMIC J-SPACE SEARCH FOR STRUCTURED LANGUAGE MODEL COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured pruning reduces language-model cost but risks removing computations needed across tasks. Capacity allocation alone leaves unresolved which equal-sized channel groups preserve useful semantic content. We propose candidate-dynamic J-space structural search to address this problem. The framework combines semantic mask characterization, joint capacity–identity optimization, and surrogate-assisted evaluation. First, a frozen Jacobian lens decomposes dense and pruned activations independently, exposing changes in sparse semantic support rather than imposing the dense model's support on every candidate. Second, a mixed-discrete representation combines component capacities and concrete masks under an exact parameter budget, while second-order compensation adjusts retained weights. Third, an online surrogate selects promising structures for full-model evaluation and updates a Pareto archive from measured behavior and semantic drift. Together, these components connect local semantic measurements to global structure selection without treating surrogate predictions as observed performance. On Qwen3.5-4B with 11.84% of parameters removed, two selected structures improve question-weighted accuracy by 1.10 and 1.19 percentage points over an equal-parameter non-J structural reference across eight benchmarks, with slightly lower WikiText-2 perplexity. The structures approach a conventional allocation reference, while results on CommonsenseQA, SciQ, and COPA reveal task-dependent trade-offs. These findings support semantic information as a useful structural-search signal, while leaving its advantage under fully matched search compute unresolved.

1 INTRODUCTION

Structured pruning reduces the cost of language models by removing coupled channels or attention groups. The difficulty lies in deciding which computations to remove while preserving performance across tasks. Unlike unstructured sparsity, structural removal changes the capacities of complete components; a locally small reconstruction error need not imply a small downstream task error. Methods that optimize both structural allocation and retained-weight compensation therefore provide an important foundation for language-model compression (An et al., 2024; Li et al., 2025).

Existing approaches address this problem from three complementary directions. Local pruning estimates the cost of removing weights or groups using activations, gradients, or second-order reconstruction. Global search distributes a fixed compression budget across components. Recovery methods adapt the retained parameters after pruning. The approaches differ in the decisions they optimize, and their contributions should be evaluated separately.

Local methods offer an economical way to construct candidate structures. SparseGPT uses approximate second-order reconstruction (Frantar & Alistarh, 2023), whereas Wanda ranks weights using their magnitudes and input activations (Sun et al., 2024). For structured removal, FLAP estimates activation fluctuations and compensates the resulting feature-map change (An et al., 2024). These measurements characterize numerical damage, but do not explicitly describe which verbalizable semantic components differ between two equal-sized masks.

Global methods account for nonuniform component sensitivity. Týr constructs local supernets and searches their sparsity allocations (Li et al., 2025). DarwinLM further evaluates candidate structures through staged recovery training (Tang et al., 2025). Minitron demonstrates the importance of combining structural pruning with distillation (Muralidharan et al., 2024). Global allocation and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

recovery are valuable, yet they leave a distinct question: can a semantic description of concrete mask identities supply useful information before expensive complete-model evaluation?

The Jacobian lens offers a vocabulary-indexed view of intermediate representations that a model is poised to verbalize, termed J-space (Gurnee et al., 2026). Its local and short-horizon nature limits the interpretation of a single global fidelity score (Yan et al., 2026). In particular, pruning can change the sparse support used to represent an activation. Comparing coefficients only on the dense model's support can therefore miss a candidate's reorganization. We use independently recomputed semantic components to characterize each mask, and test their consequences with complete-model behavior.

We introduce CJSS, a framework with three components. Candidate-dynamic J-cones describe local masks through sparse semantic decompositions. A mixed-discrete capacity–identity representation combines these masks under an exact global budget, with second-order compensation of retained weights. An online surrogate allocates full-model evaluations and maintains a Pareto archive of measured behavioral and semantic objectives. The framework makes local semantic identity available to global search while retaining an explicit distinction between predicted utility and observed performance.

Our contributions are threefold:

- We characterize local mask identities through candidate-specific sparse J-space reconstructions, allowing the support to change after pruning.
- We integrate these identities with exact-budget component allocation and surrogate-assisted full-model Pareto search. Compensation and surrogate fitting serve as established numerical tools within this framework.
- We evaluate selected structures on knowledge, commonsense, and language-modeling benchmarks, and use local interventions, multiple budgets, and recovery experiments to characterize the scope of the observed benefit.

The principal experiment uses Qwen3.5-4B and two equal-parameter structural references. Two search-selected structures improve weighted accuracy over the expanded non-J reference across 32,946 questions, but do not uniformly dominate across tasks. Additional Qwen budgets and Llama experiments probe broader settings; their experimental units and compression ratios are reported separately. Section 2 positions the method, Section 3 develops its components, and the experimental sections evaluate public task performance before examining mechanism and scale.

2 Related Work

LLM pruning. SparseGPT uses approximate second-order reconstruction for one-shot pruning (Frantar & Alistarh, 2023), while Wanda combines weight magnitude with activation magnitude (Sun et al., 2024). Structured approaches remove coupled groups that can change tensor shapes. LLM-Pruner uses gradient-based coupled-structure importance and lightweight recovery (Ma et al., 2023); FLAP combines activation fluctuation, adaptive global allocation, and bias compensation (An et al., 2024). Minitron studies pruning across width, depth, attention, and MLP axes followed by distillation (Muralidharan et al., 2024). DarwinLM makes evolutionary selection training-aware through progressively larger recovery budgets (Tang et al., 2025). Týr builds local supernets and searches global sparsity distributions with coarse-to-fine allocation (Li et al., 2025). CJSS is closest to global structural search, but expands each component variable from a capacity to a capacity–identity pair and describes identities with candidate-specific semantic measurements.

Semantic and causal representations. The J-space was introduced as a vocabulary-indexed, verbalizable workspace decoded by layerwise Jacobian lenses (Gurnee et al., 2026). J-CoT uses this coordinate system as an explicit recurrent reasoning interface (Wu et al., 2026); our method does not add recurrent inference or a carrier. Mathematical analysis of the J-lens highlights sparse paths and limited linear horizons (Yan et al., 2026). This motivates recomputing candidate supports and restricting our claim to measured masks and endpoints. Our goal is structure search, rather than matching hidden activations or claiming that a lower semantic distance guarantees better behavior.

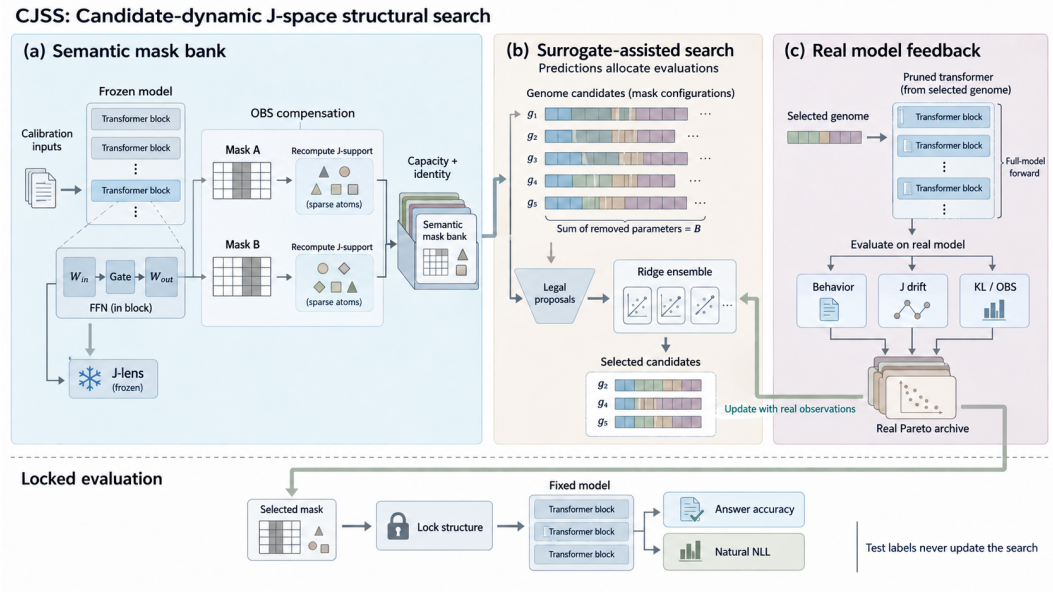

Figure 1: CJSS pipeline. Model blocks, gates, masks, and atoms are schematic. Candidate-specific J measurements describe mask identities; surrogate acquisition allocates real model evaluations, and real feedback updates the search archive. A selected structure is locked before public evaluation.

Multi-objective search and second-order compensation. The evaluated search maintains a non-dominated archive and uses surrogate-ranked mutations to allocate real evaluations. OBS supplies a mature second-order backend for compensating retained weights after removing structural groups (Hassibi et al., 1993). Neither component is claimed as a contribution; the scientific question is whether candidate-dynamic semantic identity descriptors improve the structures exposed to global search.

3 METHOD

We seek masks whose semantic differences help prioritize global structural alternatives. Figure 1 summarizes three cooperating components: semantic characterization, exact-budget structure representation, and surrogate-assisted evaluation. The following formulation distinguishes their responsibilities.

3.1 CANDIDATE-DYNAMIC SEMANTIC MASK CHARACTERIZATION

At layer ℓ, let $D_\ell \in \mathbb{R}^{V\times d}$ be the frozen Jacobian-lens direction dictionary. Starting with normalized target $\hat{h}$ and residual $r_0 = \hat{h}$, step s retrieves the highest-scoring non-excluded direction for the current residual and refits all accepted atoms:

$$
\begin{aligned}
j_s &= \arg\max_{j\notin S_{s-1}} \langle r_{s-1}, d_{\ell j}\rangle, \\
a_s^* &= \arg\min_{a\geq 0} \|\hat{h} - D_{\ell,S_{s-1}\cup\{j_s\}}^\top a\|_2^2, \\
q_s &= D_{\ell,S_{s-1}\cup\{j_s\}}^\top a_s^*, \qquad r_s = \hat{h} - q_s.
\end{aligned}
\tag{1}
$$

We accept a step only when its normalized reconstruction gain exceeds a calibration-only isotropic-direction baseline, and stop after at most 25 atoms. The dense model and each mask candidate solve Eq. 1 independently. This allows support switching and avoids comparing non-identifiable coefficients on a fixed dense support. Local descriptors include reconstructed-component change, fit defect, support occupancy, local behavior change, KL, and OBS cost. Multiple equal-cardinality

identities are retained when they expose distinct trade-offs; a conventional saliency-prefix mask is always available.

3.2 JOINT CAPACITY AND MASK-IDENTITY REPRESENTATION

Let the prunable model contain $C = 40$ ordered components. A genome is

$$g = \big((c_i, m_i)\big)_{i=1}^{C}, \tag{2}$$

where c_i is a legal number of removed structural groups and m_i identifies a concrete mask of that cardinality. If one removed group in component i contains p_i parameters, every legal genome satisfies

$$\sum_{i=1}^{C} c_i p_i = B. \tag{3}$$

We set $B = 537{,}395{,}200$. Qwen3.5-4B contains $4{,}539{,}265{,}536$ parameters in our inventory, so the budget removes 11.8388% and retains $4{,}001{,}870{,}336$ parameters. Components comprise all 32 MLP down projections and eight full-attention output projections in the hybrid architecture (Qwen Team, 2026). Attention variables operate on complete dependency groups; no genotype is repaired by silently changing B.

For each local mask, we compute a compensated retained weight using a frozen Hessian and the OBS update. A global candidate is materialized transactionally: its compensated weights are active only during its forward pass. All final metrics therefore correspond to real model execution rather than sums of local proxy scores.

3.3 SURROGATE-ASSISTED FULL-MODEL EVALUATION

Behavioral and semantic objectives. Search examples are partitioned into dependency-sensitive and lighter conditions. For candidate g, complete-model scoring yields behavior damages $L_{\mathrm{dep}}(g)$ and $L_{\mathrm{light}}(g)$. A frozen set of verbalizable causal-core directions defines a normalized reconstruction drift $D_{\mathrm{broad}}(g)$. The Pareto objectives are

$$\begin{aligned} \min_{g} \quad & \big(L_{\mathrm{dep}}(g), D_{\mathrm{broad}}(g)\big) \\ \text{s.t.} \quad & L_{\mathrm{light}}(g) \le \tau_l, \quad \mathrm{KL}(g) \le \tau_{\mathrm{KL}}, \\ & \mathrm{OBS}(g) \le \tau_{\mathrm{OBS}}, \quad \sum_i c_i p_i = B. \end{aligned} \tag{4}$$

Thresholds are frozen from the 80th percentiles of the warm real candidates. Importantly, a J score is neither substituted for complete-model behavior nor reported as downstream fitness.

Acquisition and archive updates. The candidate encoder hashes allocation, identity, local J descriptors, and selected interactions into 256 dimensions. An ensemble of four ridge regressors predicts the two objectives and constraints. Each acquisition step constructs a large legal mutation pool around the current Pareto archive and ranks it by standardized lower-confidence scores plus feasibility penalties. Only the selected batch is physically realized and evaluated. The search contains 256 warm candidates and 512 acquired candidates, for 768 unique real full-model evaluations and no failed or missing evaluations. Predicted proposals and real evaluations are counted separately.

After search, we lock: (i) the feasible minimum-dependent endpoint, (ii) the minimum-Broad endpoint, (iii) a normalized Pareto knee, and additional diversity points. Selection precedes evaluation on public benchmarks.

Overall procedure and cost. Algorithm 1 summarizes the search. Let N be the real-evaluation budget, q the batch size, M the proposal-pool size, p the feature dimension, and R the ensemble size. At most N/q rounds fit predictors and select a batch. A dense ridge fit at round t, with n_t observations, costs $O(R(n_t p^2 + p^3))$ per predicted metric; prediction costs $O(RMp)$. A straightforward pairwise Pareto filter costs $O(N^2)$, while retained feature storage scales as $O((N+M)p)$. These bookkeeping costs are additional to local bank construction and N real model evaluations, whose costs depend on input length and model size. We report measured evaluation cost separately rather than treating proposals or bank construction as free.

Algorithm 1 CJSS search procedure. All archive entries are real observations.

Input: frozen mask bank and J descriptors; exact parameter budget B; real-evaluation budget N; batch size q.
1. Generate legal warm genomes and evaluate their compensated models.
2. Freeze feasibility thresholds from warm observations; initialize the archive.
3. While the real-evaluation budget remains:
   a. Fit objective and constraint surrogates to observed genomes.
   b. Mutate archive members to form a legal capacity–identity pool.
   c. Rank proposals by confidence-aware acquisition and select q.
   d. Evaluate selected models; append measured objectives and costs.
   e. Update the feasible Pareto archive.
4. Select the minimum-dependent and trade-off roles using search scores.

Output: fixed structures for subsequent public-benchmark evaluation.

Difference from existing approaches. The semantic characterization distinguishes equal-cardinality masks through their reconstructed verbalizable content. The joint representation permits both capacity and identity to change, instead of identifying a component solely by its sparsity. Finally, global feedback comes from realized structures rather than an additive approximation to local mask losses. These are design distinctions; establishing the independent contribution of each requires matched comparisons, beyond merely obtaining a favorable search objective.

Optional recovery of selected structures. After selecting a structure, we optionally adapt its retained projections with mergeable rank-8 updates. Training combines correct-option cross-entropy and natural-text prediction with equal task-level weights. A short checkpoint uses 32,768 input tokens; continuation reaches 131,072 tokens. For each retained projection, we interpolate effective weights,

$$W_\alpha = (1-\alpha)W_{\text{short}} + \alpha W_{\text{long}}, \qquad \alpha \in \{0.2, 0.4, 0.6, 0.8\}. \tag{5}$$

Development accuracy and natural-text NLL select a coefficient before evaluating test examples. This recovery stage operates on a fixed mask and produces one model, without inference-time ensembling. The recovery experiments use the public-task-selected J-cone structures at 23.16% removal; they do not reuse the 11.84% minimum-dependent structure. The same recovery procedure is also evaluated on a generic selected structure.

4 Experimental Setup

We first test whether selected structures retain public task performance under the same parameter budget. We then examine local semantic evidence, behavior on controlled compositional tasks, multiple pruning budgets, and retained-weight recovery. Public comparisons report complete task denominators; different models and pruning ratios remain separate experiments.

4.1 Model, structures, and evaluation stages

We use the instruction-tuned Qwen3.5-4B checkpoint (Qwen Team, 2026). The official 1,000-context Qwen J-lens, frozen local Hessians, structural orders, and token bank are reused without refitting. Search uses 300 frozen cases. Before reading validation and test examples, the selector locks its roles; validation and generalization each contain 400 cases. Across the formal search, validation, and generalization stages, all requested candidates completed without technical failure.

The public endpoint experiment reconstructs four structures from their original 40-component counts, variant names, and physical mask hashes before loading the model. Evaluation uses hook-materialized compensated structures. This establishes functional endpoint quality, but it is not a standalone compact checkpoint or a deployment-speed measurement.

4.2 Knowledge, commonsense reasoning, and language modeling

We evaluate knowledge and commonsense reasoning on established benchmarks: ARC-Easy, ARC-Challenge, BoolQ, HellaSwag, OpenBookQA, RTE, and WinoGrande are evaluated zero-shot;

MMLU is evaluated five-shot (Clark et al., 2018; 2019; Zellers et al., 2019; Mihaylov et al., 2018; Sakaguchi et al., 2020; Hendrycks et al., 2021; Wang et al., 2019). There are 32,946 questions. Every model receives the same pre-rendered prompt and each complete answer option is scored by summed token log-likelihood. We report question-weighted accuracy and the unweighted mean across eight tasks.

Natural language quality is measured on the frozen WikiText-2 test stream with 296,980 next-token targets (Merity et al., 2017). A second public panel, read after structures were locked, contains official validation examples from CommonsenseQA (1,221), SciQ (1,000), and COPA (100), for 2,321 questions (Talmor et al., 2019; Welbl et al., 2017; Wang et al., 2019). CommonsenseQA, SciQ, and COPA present the answer choices in the prompt and score the likelihood of their letter labels; the primary eight benchmarks score complete option text. The eight benchmarks have been used by earlier project variants, so we distinguish evaluation after structure selection from benchmarks never inspected during the broader project.

4.3 References and comparison scope

Expanded-NonJ is an exact-budget structure frozen by a prior non-J expanded-identity selector. It shares the model, parameter budget, structural assets, and endpoint evaluator, but was not rerun with the same 768 real evaluations; it is therefore an exact-budget structural reference, not a complete equal-compute J-off ablation. **Týr-Allocation** is the all-conventional parent allocation frozen in the formal registry, again at the same exact budget. It is not the separate 23.16%-removal released-source Qwen adaptation used in a later study. The dense reference uses the same model and prompt protocol and was already evaluated completely; we do not spend compute repeating it.

Controlled compositional reasoning tasks. For search and internal validation, we generate five task families: branching checksums, indexed substitutions, modular affine arithmetic, precedence-constrained scheduling, and toggle-state tracking. Each underlying instance yields a dependency-sensitive question requiring intermediate computation and a paired factual-retrieval question. Calibration, search, validation, and test prompt-generalization splits contain 30, 300, 400, and 400 examples, respectively, balanced between conditions. Prompt and entity reservoirs are disjoint; task families are shared. These are controlled compositional tasks, distinct from the public benchmarks above. We measure the change in the model's answer-preference score and semantic reconstruction, rather than exact-match chain-of-thought trajectories.

4.4 Statistics

For paired multiple-choice comparisons, b counts questions correct only for the left model and c questions correct only for the right model. We report the exact two-sided McNemar binomial test on $b + c$. The two CJSS endpoint roles were locked before the public endpoint scores and are both reported; we do not choose between them using public test accuracy.

5 Results

5.1 Knowledge and commonsense benchmarks

Table 1 is the main result. Both locked CJSS roles improve weighted benchmark accuracy over the exact-budget non-J reference by more than one percentage point while slightly reducing perplexity. Against the non-J reference, CJSS has 2,362 left-only and 2,000 right-only correct answers ($p = 4.53 \times 10^{-8}$); CJSS (Pareto knee) has 2,369 and 1,978 ($p = 3.24 \times 10^{-9}$). The two roles are within 29 questions of each other, providing evidence that the result is not confined to one selected extreme.

The two CJSS structures are within 0.15 percentage points of the conventional allocation reference (61.06%), while improving over Expanded-NonJ by 1.10 and 1.19 points. Table 1 includes the uncompressed model and perplexity for all structures.

Table 1: Qwen3.5-4B endpoint results at an exact removal budget of 537.4M parameters (11.84%). Higher is better except perplexity. Dense is an unpruned reference.

Structure	Correct	Weighted	Macro	WikiText-2 PPL↓
Dense	23,406	71.04	75.88	8.671
Týr-Allocation	**20,116**	**61.06**	**67.70**	**10.955**
CJSS (Pareto knee)	20,098	61.00	66.80	11.156
CJSS	20,069	60.91	66.88	11.154
Expanded-NonJ	19,707	59.82	67.07	11.204

Table 2: Per-task accuracy (%) on the eight public benchmarks. The non-J reference has a higher macro-average, while CJSS gains substantially on BoolQ and MMLU.

Structure	ARC-C	ARC-E	BoolQ	HellaSwag	MMLU	OpenBookQA	RTE	WinoGrande
CJSS	81.23	92.30	**79.05**	44.25	62.01	73.80	48.38	53.99
CJSS (Pareto knee)	81.06	92.09	**79.14**	44.41	**62.13**	73.60	47.65	54.30
Expanded-NonJ	**82.42**	**93.27**	66.97	**45.34**	60.92	**74.80**	**58.12**	**54.70**

5.2 Task specialization and transfer to CommonsenseQA, SciQ, and COPA

Table 2 exposes the source of the weighted gain. CJSS improves BoolQ by roughly 12 points and MMLU by 1.1–1.2 points relative to Expanded-NonJ, whereas the non-J reference is better on the other six tasks. Since MMLU and HellaSwag together account for 73% of all evaluated questions, weighted and macro accuracy answer different questions.

On CommonsenseQA, SciQ, and COPA combined, CJSS, CJSS (Pareto knee), Expanded-NonJ, Týr-Allocation, and dense obtain 1,866, 1,853, 1,933, 1,851, and 2,018 correct answers, respectively. Expanded-NonJ exceeds the two CJSS roles by 67 and 80 questions ($p = 2.60 \times 10^{-5}$ and 8.78×10^{-7}). These scores characterize the task dependence of the aggregate improvement.

5.3 Causal module identification and local mask selection

The full-model experiment builds on two earlier positive stages rather than introducing J-space only at the final search. First, a three-seed intervention screen confirmed 25 of 64 modules under zero, corruption, patch, rescue, and family-matched controls (7 Gated DeltaNet, 4 full-attention, and 14 MLP modules). Second, the candidate-dynamic sparse J-cone protocol was tested on four local FFN layers and 400 fresh test cases. Its pooled score exceeded a matched Týr local selector by 0.003281, with a paired 95% interval $[0.001016, 0.005547]$; two of four layers were positive, and layer 19 also exceeded a geometry-matched rotated control by 0.0100. These results motivate semantic identity features but do not themselves establish full-model accuracy. Table 3 details the local experiment. Appendix B gives the complete budget and cross-architecture comparisons.

5.4 Compositional reasoning under structural compression

Table 4 reports the frozen semantic/behavioral evaluation that originally promoted the CJSS endpoints. Lower is better. The minimum-dependent role improves the targeted dependent and Broad metrics on both splits relative to the canonical allocation. These measurements motivate endpoint evaluation; they do not substitute for it.

Larger independent prompt confirmation. The same frozen primary CJSS structure was also evaluated on two additional 1,600-example compositional-task splits. Dependent damage is $-0.13578/-0.16063$, compared with $-0.00844/+0.00031$ for the conventional allocation. Semantic drift is $0.00995/0.01089$, compared with $0.02180/0.02172$. All nine registered comparison structures complete both splits. These additional cases confirm the internal-score direction using new prompts; they are separate from the three-budget search and do not increase its training denominator.

Table 3: Local candidate-dynamic J-cone selection on Qwen. Entries are continuous behavioral-score differences, not accuracy percentage points. The pooled evaluation covers four FFN layers and 400 test cases.

Comparison	J advantage	95% interval
Pooled J vs. conventional selector	$+0.003281$	$[+0.001016, +0.005547]$
Layer 19 vs. conventional selector	$+0.010000$	—
Layer 19 vs. rotated dictionary	$+0.010000$	—
Layer 22 vs. conventional selector	$+0.003125$	—
Layer 22 vs. rotated dictionary	0	—

Table 4: Internal validation and test measurements for the locked minimum-dependent role and canonical Týr allocation. Lower is better.

Split	Structure	Dependent	Light	Broad drift
Validation	CJSS	-0.1106	-0.5994	0.00992
	Týr-Allocation	-0.0572	$+0.1975$	0.02184
Generalization	CJSS	-0.0981	-0.4578	0.01085
	Týr-Allocation	-0.0481	$+0.0250$	0.02145

5.5 Scaling across pruning budgets

We additionally evaluate the budget-conditioned CJSS-Scale search on Qwen3.5-4B at nominal pruning budgets of 12.5%, 25%, and 37.5%. This is a separate three-budget search using shared structural assets, rather than relabeling the primary 11.84% endpoint run. Each budget receives 1,280 real full-model evaluations (256 warm and 1,024 acquired), totaling 3,840 unique structures. Four roles per budget are selected before evaluating 1,000 validation and 1,000 unseen compositional-task examples from the controlled compositional tasks.

At 12.5% and 25%, the minimum-dependent structures improve the measured dependent score relative to the dense reference on both splits; at 25%, the damage is $-0.35013/-0.30738$. These are improvements on the compositional-task score, not public benchmark accuracy. The 37.5% setting exposes a more difficult compression regime. Appendix B reports all 12 pre-selected structures so that budget-level conclusions do not depend on post-hoc role selection. This study measures controlled behavioral and semantic preservation across budgets; the public results in Table 1 are from the separately locked primary experiment.

Cross-architecture evidence. On Llama-3.1-8B-Instruct at nominal 25% removal, four locked roles complete two 1,000-example compositional-task splits. Relative to the same-runtime pruning reference, the minimum-dependent role improves factual-retrieval damage by 0.9491 on validation (95% CI $[0.6687, 1.2328]$) and 0.9168 on unseen prompts ($[0.6626, 1.1778]$). The minimum-semantic and knee roles reduce Broad drift on both splits. Public MMLU accuracy is 53.77% for the initial comparison and 57.59% after search extension, versus 51.92% for the reference. Full task measurements are given in Appendix B.4; these knowledge-task gains accompany a broader quality trade-off.

5.6 Improving compressed models through recovery

At 23.16% parameter removal, recovery improves the J-cone-selected structure from 44.70% to 48.08% and 48.70% weighted accuracy in two training seeds, while reducing WikiText-2 perplexity from 21.07 to 17.45 and 17.29 (Table 8). The structures and interpolation coefficients are fixed before testing. Both recovered models also improve over the unrecovered pruning reference on CommonsenseQA, SciQ, and COPA individually (Appendix C). These results extend the structure-search framework with a practical retained-weight adaptation stage.

The recovered second-seed model is physically exported from 4.539B to 3.488B parameters. Resident allocated memory falls from 9.08 to 7.00 GB and measured 128-token time to first token falls from 51.21 to 46.58 ms; decode throughput remains approximately 36.3 tokens/s. The generic recovery

control establishes that recovery is useful across structures; the first J-origin seed differs from it by 41 answers ($p = 0.475$), so the recovery gain is not attributed exclusively to J-space.

6 DISCUSSION

From semantic identities to executable structures. The intervention screen, local mask evaluation, and global search answer complementary questions. Module interventions establish functional sensitivity; local candidate comparisons test whether semantic re-decomposition supplies a useful selection signal; the global experiment evaluates combinations under an exact parameter constraint. The local J-cone advantage and the full-model test improvements motivate retaining mask identity as a search variable alongside capacity. These studies have different experimental units and are not pooled into one effect size.

Consistency across locked roles. CJSS and CJSS (Pareto knee) improve weighted benchmark accuracy over Expanded-NonJ by 362 and 391 correct answers, respectively, with similar perplexity. They are two roles from one search, not independent search seeds. Their agreement indicates that the endpoint gain is not confined to the minimum-dependent extreme of the archive. The per-task results locate the public advantage primarily in BoolQ and MMLU.

Scope of the comparison. The non-J reference matches the parameter budget, but not the complete search-compute allocation. The conventional allocation remains competitive, and the second public panel favors Expanded-NonJ. These comparisons establish a task-dependent structural improvement rather than equal-compute J-specific superiority. The primary public comparison uses Qwen at one exact budget; the additional Qwen study covers three nominal budgets on controlled tasks, and Llama-3.1-8B-Instruct contributes a separate 25% public evaluation. These studies broaden model and budget coverage without forming a fully crossed model-by-budget public-benchmark grid. Physical deployment measurements apply to the recovered 23.16% model; the 11.84% structures are evaluated with the compensated runtime.

7 CONCLUSION

We presented CJSS for structured language-model compression under an exact parameter budget. Candidate-dynamic J-cones characterize semantic changes caused by local masks; a joint capacity–identity representation exposes those alternatives to global optimization; and surrogate-assisted evaluation uses complete-model observations to select structures. Their combination connects verbalizable representations to concrete pruning decisions.

On Qwen3.5-4B at 11.84% parameter removal, two selected structures improve weighted accuracy over an equal-parameter expanded non-J reference by 1.10 and 1.19 percentage points across eight benchmarks, with slightly lower WikiText-2 perplexity. Performance remains close to the conventional allocation reference and varies across task distributions. Separate budget, Llama, and recovery experiments characterize broader behavior without establishing a fully matched end-to-end J-specific advantage.

Future work should prioritize equal-compute semantic ablations, structure selection that balances multiple task distributions, and physical deployment of the selected architectures. These directions test whether the observed task-specific benefits extend to a more general compression advantage.

REPRODUCIBILITY STATEMENT

Appendix A specifies the model inventory, component granularity, exact budget, search counts, data splits, endpoint tasks, and statistical tests. Candidate definitions, selected structures, per-question predictions, and search budgets are retained for reproducibility. The anonymized supplemental code will include the candidate registry, search configuration, endpoint evaluator, and result manifests; frozen model weights and public datasets remain governed by their original licenses.

Ethics Statement

This work studies compression of an open-weight language model using public NLP benchmarks. It introduces no human-subject data and makes no claims about consciousness from the global-workspace analogy. Compression can change behavior unevenly across task families; we therefore report task-level reversals rather than presenting the aggregate gain as universal reliability.

AI Use Statement

Generative AI tools were used for code assistance, experiment monitoring, language editing, and preparation of the manuscript draft. The authors selected the research questions, executed and audited the experiments, verified numerical claims against saved artifacts, and remain responsible for all content and citations.

References

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Fluctuation-based adaptive structured pruning for large language models. In *AAAI Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2312.11983.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2301.00774.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models, 2026. URL https://arxiv.org/abs/2607.15495.

Babak Hassibi, David G. Stork, and Gregory J. Wolff. Optimal brain surgeon: Extensions and performance comparisons. In *Advances in Neural Information Processing Systems*, 1993. URL https://proceedings.neurips.cc/paper/1993/hash/b056eb1587586b71e2da9acfe4fbd19e-Abstract.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Guanchen Li, Yixing Xu, Zeping Li, Ji Liu, Xuanwu Yin, Dong Li, and Emad Barsoum. Týr-the-pruner: Unlocking accurate 50% structural pruning for llms via global sparsity distribution optimization. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.09657.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models, 2023. URL https://arxiv.org/abs/2305.11627.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation, 2024. URL https://arxiv.org/abs/2407.14679.

Qwen Team. Qwen3.5-4b model card. https://huggingface.co/Qwen/Qwen3.5-4B, 2026. Accessed September 2026.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale, 2020. URL https://arxiv.org/abs/1907.10641.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.11695.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Shengkun Tang, Oliver Sieberling, Eldar Kurtic, Zhiqiang Shen, and Dan Alistarh. Darwinlm: Evolutionary structured pruning of large language models, 2025. URL https://arxiv.org/abs/2502.07780.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems, 2019. URL https://arxiv.org/abs/1905.00537.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Junde Wu, Jiayuan Zhu, Fengling Liu, Minhao Hu, and Jiazhen Pan. J-cot: Chain-of-thought in j-space, 2026. URL https://arxiv.org/abs/2607.21981.

Shi-Qi Yan, Kai-Xuan Ding, Chao-Hong Tan, Qian Chen, Wen Wang, Xiangang Li, and Zhen-Hua Ling. Short horizons and sparse concepts: A mathematical view of the readout in the j-lens, 2026. URL https://arxiv.org/abs/2608.25347.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

A Detailed Experimental Protocol

A.1 Model and component inventory

The evaluated checkpoint contains 4,539,265,536 parameters. The exact removal budget is 537,395,200 parameters, leaving 4,001,870,336 parameters. The 40 variables are ordered as all 32 MLP down projections plus the full-attention output projections at layers 3, 7, 11, 15, 19, 23, 27, and 31. FFN variables remove registered groups of 16 input channels; attention variables remove complete dependency groups. Every role is reconstructed from $(c_i, m_i)_{i=1}^{40}$ and checked against the physical mask identity saved when the selector was locked.

A.2 Search accounting

The formal search evaluates 256 warm candidates and 512 surrogate-acquired candidates, totaling 768 unique real physical genotypes and 36,555,264 model-input tokens. Real-evaluation wall time is 5,654.2 seconds. The frozen warm-Q25 threshold is first reached by evaluation 25 (wall-time upper bound 230.8 seconds); best dependent damage improves from 0.11 after 16 evaluations to -0.21375 after 768, while best Broad drift improves from 0.01747 to 0.00715. The cheap proposal pool, predictions, and real evaluations are separate ledgers. Six Pareto candidates are retained and five unique candidate roles are locked before validation and test examples. Validation and generalization each evaluate 400 cases for every requested logical and unique physical role. The run contains zero failed or missing scientific evaluations.

A.3 Sparse solver, local identities, and surrogate features

For each residual, vocabulary retrieval first forms a BF16 top-8,192 shortlist and then reranks the top 512 directions in FP32. Accepted directions are unit-normalized. The nonnegative least-squares refit in Eq. 1 uses projected gradient descent for at most 64 iterations with tolerance 10^{-7}; pursuit stops at 25 atoms, no-positive-atom, or when marginal gain fails to exceed the median isotropic null calibrated with two frozen seeds. For every one of the 93 addressable $(\text{component}, \text{count})$ coordinates, 16 deterministic equal-cardinality masks are measured and at most six nonconventional identities are locked; adding the conventional prefix gives exactly seven identities per coordinate.

The surrogate encoder is a 256-dimensional signed SHA-256 feature hash. It includes normalized count per component, one token for every concrete (i, c_i, m_i) identity, five local descriptors (OBS–KL compromise l_{comp}, J-cone loss, occupancy drift, component mass, and fit defect), and pairwise tokens between active nonconventional identities. Four bootstrap ridge models with $\lambda = 10^{-3}$ provide means and standard deviations. Acquisition uses mean minus 0.5 standard deviations for the two minimized objectives, mean plus one standard deviation for feasibility constraints, and a fixed penalty of 100 for constraint violation. Each round ranks 30,000 legal proposals and physically evaluates 16.

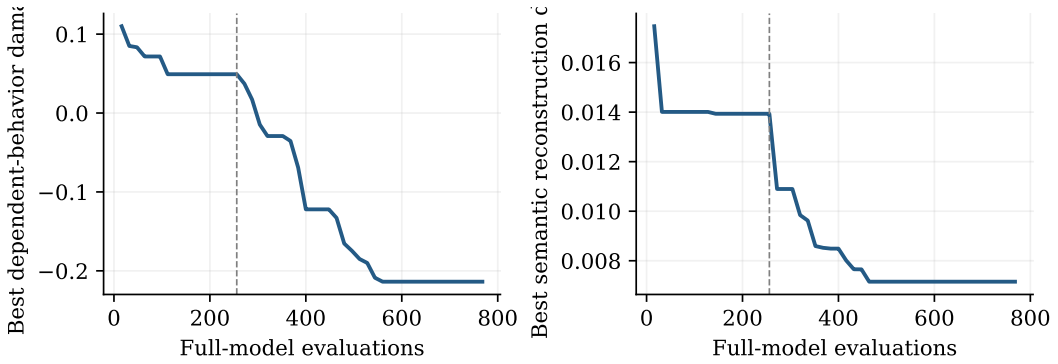

Figure 2: Best observed objectives across all 768 real full-model evaluations. The dashed line marks the 256-candidate initialization. Each curve follows its own best candidate; the curves do not imply that one structure attains both minima.

Table 5: All five pre-selected CJSS structures on the constructed compositional reasoning tasks (400 examples per split). D: dependent-behavior damage; L: factual-retrieval damage; J: semantic reconstruction drift. Lower is better.

	Validation			Unseen prompts/entities		
Structure	D	L	J	D	L	J
CJSS	-.11063	-.59938	.00992	-.09813	-.45781	.01085
Minimum semantic drift	-.09938	-.57188	.00854	-.05969	-.62063	.00790
CJSS (Pareto knee)	-.09438	-.59750	.00895	-.08250	-.54500	.00898
Pareto structure 4	-.08000	-.56750	.00870	-.07969	-.54375	.00860
Pareto structure 5	-.07375	-.60500	.00903	-.08156	-.54188	.00903
Conventional allocation	-.05719	+.19750	.02184	-.04813	+.02500	.02145

A.4 Benchmark datasets and evaluation protocols

Table 6: Public endpoint composition.

Endpoint	Cases	Protocol
ARC-Easy	2,376	0-shot option likelihood
ARC-Challenge	1,172	0-shot option likelihood
BoolQ	3,270	0-shot option likelihood
HellaSwag	10,042	0-shot option likelihood
OpenBookQA	500	0-shot option likelihood
RTE	277	0-shot option likelihood
WinoGrande	1,267	0-shot option likelihood
MMLU	14,042	5-shot option likelihood
Eight-task total	32,946	weighted and task macro
CommonsenseQA, SciQ, COPA	2,321	answer-label likelihood
WikiText-2	296,980 tokens	next-token perplexity

A.5 Endpoint role provenance

CJSS minimizes dependent behavior damage within the frozen feasible archive. CJSS (Pareto knee) minimizes the sum of normalized dependent and Broad objectives on the Pareto front. Expanded-NonJ is a frozen exact-budget structure from an earlier non-J expanded-identity selector; its source search is not counted as a rerun of the CJSS optimizer. Týr-Allocation is the all-conventional frozen parent allocation under the same exact budget. Dense is evaluated without candidate hooks.

A.6 Benchmark visualization

A.7 Statistical counts

For CJSS versus Expanded-NonJ on the eight public benchmarks, both are correct on 17,707 questions, only CJSS is correct on 2,362, only non-J is correct on 2,000, and both are wrong on 10,877. For CJSS (Pareto knee) versus Expanded-NonJ, the corresponding discordant counts are 2,369 and 1,978. On CommonsenseQA, SciQ, and COPA, the discordant counts are 91 versus 158 and 91 versus 171, respectively. Exact two-sided McNemar tests use only these discordant pairs.

B Multi-budget and Cross-Architecture Evaluation

B.1 Exact parameter budgets

Nominal pruning rates in the scale experiments refer to the eligible structural parameter inventory. Table 9 gives both that inventory and the exact removed count. The earlier primary CJSS experiment has its own 537,395,200-parameter constraint and is not the same structure or budget as the nominal 12.5% Scale experiment.

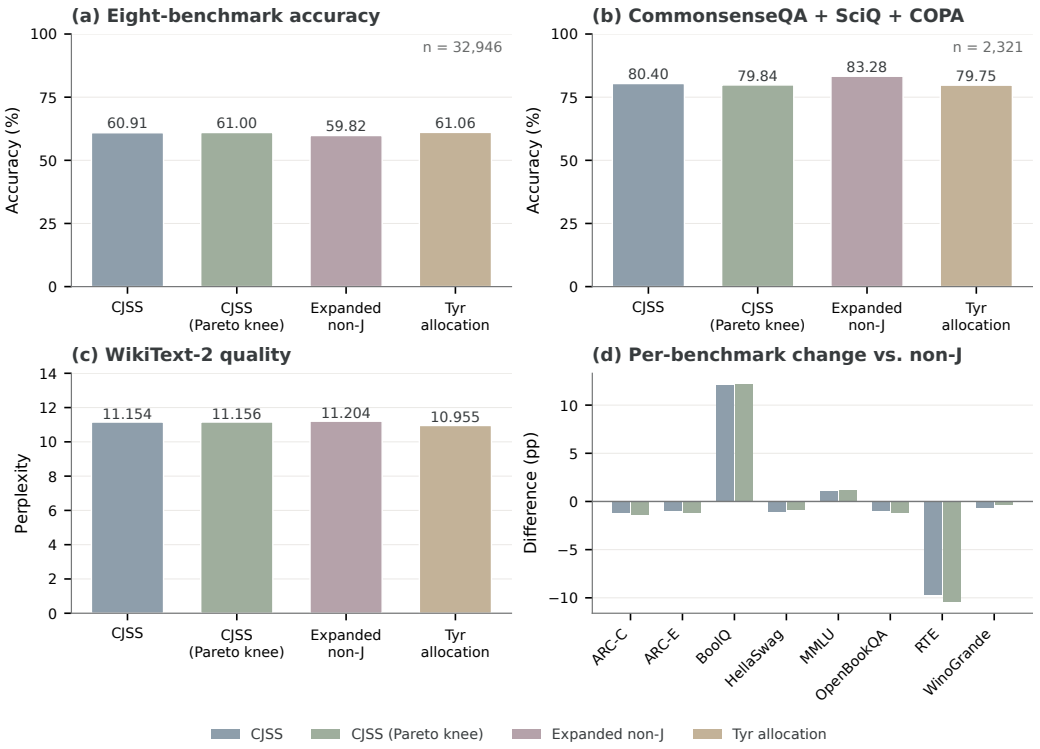

Figure 3: Qwen3.5-4B at 11.84% parameter removal. (a) Question-weighted accuracy across ARC-Easy, ARC-Challenge, BoolQ, HellaSwag, OpenBookQA, RTE, WinoGrande, and MMLU (32,946 questions). (b) Accuracy on CommonsenseQA, SciQ, and COPA (2,321 questions). (c) WikiText-2 perplexity. (d) Per-benchmark accuracy differences against Expanded-NonJ. CJSS is the minimum-dependent-damage solution; CJSS (Pareto knee) is the trade-off solution from the same search. Tyr allocation denotes the conventional allocation reference.

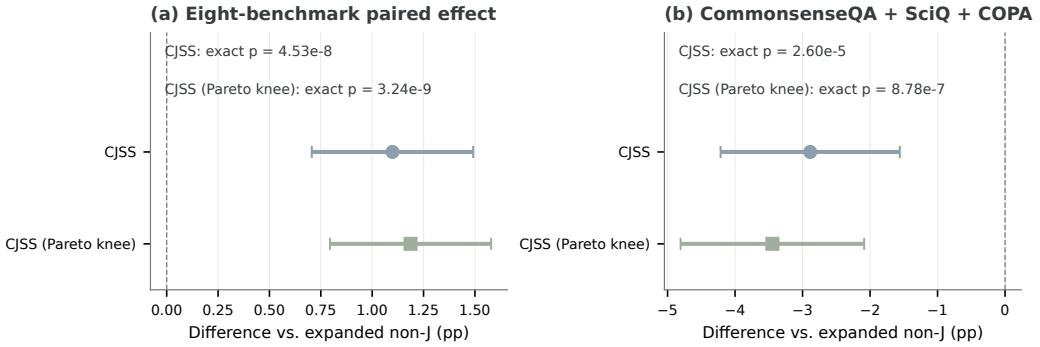

Figure 4: Paired accuracy contrasts against Expanded-NonJ. Error bars are approximate 95% normal confidence intervals for paired binary differences, computed from the discordant question counts; they are not standard deviations across independent searches. Exact McNemar p-values are shown separately.

Table 7: Qwen3.5-4B across three nominal pruning budgets. Rows track the search-selected minimum-dependent role, not the best role chosen on each test split. D: dependent-behavior damage; J: Broad semantic drift (lower is better).

Nominal budget	Real evaluations	Val. D	Test D	Val. J	Test J
12.5%	1,280	-0.21550	-0.12762	0.01041	0.01115
25%	1,280	-0.35013	-0.30738	0.01826	0.02136
37.5%	1,280	0.02013	0.16462	0.05151	0.04902

Table 8: Qwen3.5-4B recovery experiments at 23.16% removal. Accuracy aggregates the eight knowledge and commonsense benchmarks. J-cone and generic structures receive the same recovery procedure.

Structure and training	Correct	Accuracy (%)	Macro (%)	PPL
Pruning reference, unrecovered	14,432	43.81	51.01	14.2515
J-cone-selected, unrecovered	14,727	44.70	51.18	21.0663
J-cone-selected + recovery, seed 1	15,841	48.08	56.77	17.4503
J-cone-selected + recovery, seed 2	16,045	48.70	57.97	17.2906
Generic selected + recovery	15,800	47.96	55.89	17.6796

Table 9: Exact compression budgets verified from the original remote registries. Whole-checkpoint rates for Qwen use 4,539,265,536 parameters; Llama's rate is reported against its eligible inventory.

Model/study	Nominal	Eligible parameters	Removed parameters	Whole Qwen (%)
Qwen primary	—	—	537,395,200	11.8388
Qwen Scale	12.5%	4,205,751,296	525,598,720	11.5789
Qwen Scale	25%	4,205,751,296	1,051,361,280	23.1615
Qwen Scale	37.5%	4,205,751,296	1,577,123,840	34.7440
Llama-3.1-8B	25%	8,030,261,248	2,007,498,752	—

B.2 QWEN3.5-4B: ALL SELECTED STRUCTURES

The three nominal budgets use the budget-conditioned search's own parameter ledger; their names must not be substituted for exact whole-checkpoint percentages. Each budget contributes four search-selected structures. All 12 complete both 1,000-example splits, with no failed or missing evaluations. The nominal 12.5% row in this experiment is not the earlier 537,395,200-parameter CJSS winner. The combined search consumed approximately 182.37M model-input tokens; its recorded evaluation wall time is 16,802.5 seconds, excluding the separately constructed local bank.

Table 10: Complete three-budget Qwen test results. D: dependent damage; L: factual-retrieval damage; J: semantic drift. Each row is one pre-selected structure; all measures are lower-is-better.

		Validation			Unseen prompts/entities		
Budget	Role	D	L	J	D	L	J
12.5%	Min. dependent	-0.2155	-0.3728	0.0104	-0.1276	-0.3170	0.0111
12.5%	Min. semantic drift	-0.1343	-0.9115	0.0069	-0.1174	-1.0736	0.0071
12.5%	Pareto knee	-0.2116	-0.3779	0.0094	-0.1340	-0.3000	0.0101
12.5%	Diversity	-0.1605	-0.5414	0.0101	-0.1491	-0.4223	0.0102
25%	Min. dependent	-0.3501	-1.7123	0.0183	-0.3074	-1.6905	0.0214
25%	Min. semantic drift	-0.2911	-1.9083	0.0122	-0.2106	-2.0294	0.0132
25%	Pareto knee	-0.3134	-1.6540	0.0175	-0.2839	-1.5508	0.0209
25%	Diversity	-0.2756	-0.7541	0.0146	-0.1820	-0.9661	0.0169
37.5%	Min. dependent	0.0201	1.0382	0.0515	0.1646	1.6398	0.0490
37.5%	Min. semantic drift	0.0399	1.4300	0.0267	0.0915	1.7091	0.0317
37.5%	Pareto knee	0.0196	0.8811	0.0381	0.1494	1.3707	0.0396
37.5%	Diversity	-0.0177	0.6119	0.0352	0.1046	0.9650	0.0368

B.3 LLAMA: PRUNING BUDGET VERSUS SEARCH BUDGET

The Llama-3.1-8B-Instruct registry and final search artifact specify one nominal 25% pruning budget: exactly 2,007,498,752 parameters out of an 8,030,261,248-parameter eligible inventory (24.99917%). Its original search evaluates 1,280 unique structures. A subsequent study evaluates expanded-search and cross-family-exchange branches at 1,472 observations each and a single random-search branch at 432 observations. These are the evaluation counts at the public-score lock, not three pruning ratios. Later saved progress reaches 1,568 observations for each of expanded and exchange search and 512 for random search; public scores below remain attached to the earlier 1,472/432 locks, rather than being reassigned to those later checkpoints. Table 11 reports all six selected roles from that extension.

Table 11: Llama-3.1-8B-Instruct at the same nominal 25% pruning ratio across search branches and evaluation budgets. Search counts denote completed observations; the planned 5,120-observation extension was not completed.

Branch	Selected role	Search count	Weighted (%)	Macro (%)	PPL
Expanded search	Min. dependent	1,472	57.68	64.00	31.847
Expanded search	Min. KL	1,472	55.51	61.77	25.272
Exchange	Min. dependent	1,472	56.46	61.43	29.564
Exchange	Min. KL	1,472	55.51	61.77	25.272
Random	Min. dependent	432	50.19	53.28	28.928
Random	Min. KL	432	50.02	55.87	26.616

B.4 LLAMA: COMPLETE TEST AND PUBLIC-TASK COMPARISONS

The main comparison below uses the retained same-runtime CJSS/Týr evaluation. The initial CJSS and reference rows consistently use the same comparison evaluation. Search-extension rows use their pre-selected structures and are shown separately.

Table 12: Llama-3.1-8B-Instruct, nominal 25% removal: all four search-selected roles on constructed compositional tasks. D: dependent damage; L: factual-retrieval damage; J: semantic drift; lower is better.

	Validation			Unseen prompts		
Role	D	L	J	D	L	J
Min. dependent	-0.4045	-2.2132	0.0561	-0.3096	-1.7803	0.0575
Min. semantic	-0.3244	-2.4830	0.0425	-0.2044	-1.6553	0.0416
Pareto knee	-0.3974	-2.4038	0.0461	-0.2439	-1.5802	0.0445
Diversity	-0.3550	-2.1749	0.0521	-0.2069	-1.5084	0.0542
Pruning reference	-0.1736	-1.2641	0.0478	-0.0319	-0.8635	0.0491

Table 13: Full public evaluation on Llama-3.1-8B-Instruct, 25% nominal removal. Initial CJSS and the reference use the saved same-runtime comparison; expanded search is the 1,472-observation locked structure.

Dataset	Examples	Initial CJSS	Expanded search	Reference
ARC-Easy	2,376	83.92	88.47	87.46
ARC-Challenge	1,172	66.55	72.87	73.38
BoolQ	3,270	80.52	71.68	82.48
HellaSwag	10,042	43.06	43.64	48.53
OpenBookQA	500	67.80	71.60	72.00
RTE	277	47.65	47.29	48.74
WinoGrande	1,267	58.80	58.88	60.30
MMLU	14,042	53.77	57.59	51.92

C Recovery Configuration and Transfer Evaluation

The recovery structure is selected from the J-cone identity bank using the public training/development split. It removes 1,051,361,280 parameters. The AdamW learning rate is 2×10^{-4}, weight decay 0.01, gradient norm limit 1.0, and low-rank scale 16. Each step contains 4,096 input tokens. The first eight examples per task (64 total) train the short checkpoint; continuation uses the remaining 192 training examples. The 256-example development split chooses the effective-weight interpolation coefficient; neither public test questions nor their labels enter gradient updates or coefficient selection.

Table 14: Correct answers after recovery on the official validation sets of CommonsenseQA, SciQ, and COPA. Options appear in the prompt and are scored by answer-letter likelihood.

Model	CommonsenseQA 1,221	SciQ 1,000	COPA 100	Total 2,321
Dense	965	955	98	2,018
Pruning reference, unrecovered	625	677	69	1,371
J-cone-selected + recovery, seed 1	664	753	78	1,495
J-cone-selected + recovery, seed 2	677	766	78	1,521
Generic selected + recovery	674	767	76	1,517

The exported second-seed model reloads without dense weight shards. Its predictions agree with compensated-runtime predictions on 191 of 192 checked examples. The measured deployment benefit is parameter and memory reduction and lower time to first token, with unchanged decode throughput. All these values apply to the recovered 3.488B checkpoint.